Exact set-based hypothesis testing of correlated normal statistics using a distributional equivalence in a high-dimensional space

Noah Lorincz-Comi, PhD
lorincn at ccf.org
03/10/2025

Abstract

Gene-based association test statistics based on the sum of chi-square statistics from surrounding SNPs in linkage disequilibrium (LD) have no known null distribution. As a result, researchers test gene based null hypotheses using simulation and permutation approaches which have limited theoretical support. We present an exact test of the gene-based null hypothesis which does not require specifying the exact null distribution of the gene-based association test statistic. This test is based on a known distribution equivalence between the gene-based test statistic and a weighted sum of independent chi-square statistics. We leverage the joint distribution of the latter quantity to define a critical region which leads to exactly controlled Type I error in gene-based association testing.

1 Gene-based association testing

Individual SNP effect sizes are generally weak such that detecting their association with a phenotype using genome-wide association studies (GWAS) may be unlikely in most cases. Researchers have therefore considered leveraging information across multiple SNPs in and around the body of a gene to aggregate effect sizes and have a more powerful test. We refer to this type of test as a ‘gene-based association test’, which may be more powerful than standard SNP-based tests because of the reduced multiple testing burden when applied genome-wide, and because multiple SNPs with weak effects can still lead to a powerful gene-based association test. The most common statistic to test the gene based null hypothesis, which in words is typically that no SNP in the gene-specific set is associated with the phenotype, is the sum of SNP-specific chi-square statistics. These chi-square statistics are the squared Z-statistics from the regression of the outcome phenotype on the genotypes of each SNP, generally from models fitted separately for each SNP. We introduce this statistic for a single gene in the following.

Let G_{ij} represent the standardized dosage-coded genotype of the jth SNP for the ith person in a GWAS, and Y_i represent the value of the unit-variance outcome phenotype, assumed for simplicity for now to be continuous and that $E(G_{ij}) = 0$ and $Var(G_{ij}) = 1$. Let $\mathcal{G} = \{j\text{: SNP } j \text{ in vicinity of gene}\}$ be the set of m SNPs in the vicinity of the select gene, generally chosen such that the distance between the SNP and transcription start site (TSS) of the gene is less than 50K base pairs. The GWAS-estimated effect of G_{ij} on Y_i from n independent subjects is $\hat{\beta}_j = n^{-1} \sum_i G_{ij} Y_i$ and its estimated standard error be approximately $n^{-1/2}$. Since each SNP is tested separately, the correlation between $\hat{\beta}_j$ and $\hat{\beta}_t$ is approximately r_{jt}, the LD correlation between the jth and tth SNPs. Let $\mathbf{R} = (r_{jt})$ represent the LD matrix for SNPs in $\mathcal{G}$ with positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. The statistic $Z_j = n^{-1/2} \hat{\beta}_j$ can be used to test $H_{0j}\text{:}\, \beta_j = 0$ vs $H_{1j}\text{:}\, \beta_j \neq 0$, and the standard gene-based test statistic is $T = \sum_{\mathcal{G}} Z_j^2$, which is used to test $H_0\text{: } \cap_{\mathcal{G}}\, \beta_j = 0$ vs $H_1\text{: } \cup_{\mathcal{G}}\, \beta_j \neq 0$. The challenge is that the distribution of T under H_0, which we will denote for simplicity as $\gamma(T)$, does not have a closed-form expression. Researchers have generally resorted to permutation or simulation to approximate $\gamma(T)$, but recently Lorincz-Comi et al. (2025a) proposed to approximate $\gamma(T)$ with $\Gamma(\alpha,\theta)$ where $\theta = 2\text{trace}(\mathbf{RR})/m$ and $\alpha = m/\theta$ using the shape-scale parameterization. This parameterization is accomplished by matching the first two moments of T to that of a Gamma distribution. In the next section, we will show that $\gamma(\cdot)$ is explicitly not equivalent to $\Gamma(\alpha,\theta)$, and that no Gamma distribution can exactly match $\gamma(\cdot)$.

2 Approximation of moment-matched Gamma distribution

We begin by noting that $\mathbf{z} = (Z_j)$ has distribution $N(\mathbf{0},\mathbf{R})$ under H_0 and that $S = \sum_{j=1}^{m} \lambda_j Y_j$ has the same distribution as $\mathbf{z}^\top\mathbf{z}$ where $Y_j = \mathbf{u}_j^\top \mathbf{z}$, $\mathbf{R} = \mathbf{U\Lambda U}^\top = \sum_{j=1}^{m} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top$. This property states that the gene-based test statistic $T = \mathbf{z}^\top\mathbf{z}$ is distributionally equivalent to a weighted sum of independent 1df chi-squares whose weights are the eigenvalues of the LD matrix to which the SNPs in $\mathcal{G}$ correspond. In this section, we show that the cumulant function of S is not that of the distribution $\Gamma(\alpha,\theta)$, suggesting that the Gamma distribution is not the exact distribution of T. If $T \sim \Gamma(\alpha,\theta)$, then the cumulant function of T under the Gamma distribution is

$$C_\Gamma(t) = -\alpha \log(1 - t\theta)\,, \qquad t < \theta\,.$$

Similarly, if $Y_j \sim \chi^2(1)$, the cumulant function of S is

$$C_\chi(t) = -\frac{1}{2}\sum_{j=1}^{m} \log(1 - 2t\lambda_j)\,, \qquad t < \frac{1}{2\lambda_j}\,.$$

That is, $C_\chi(t)$ is the cumulant function of the true distribution of T under H_0, $\gamma(\cdot)$, and $C_\Gamma(t)$ is the cumulant function of the approximation of $\gamma(\cdot)$ by $\Gamma(\alpha,\theta)$. We now show that the first two moments of $\Gamma(\alpha,\theta)$ match those of $\gamma(\cdot)$, but that the third and fourth order moments do not. The first two moments of $\Gamma(\alpha,\theta)$ are made to match those of $\gamma(\cdot)$ by the parameterization using (α,θ), so we will omit them in the following. By differentiating each cumulant function, we find for skewness $\kappa(3)$ that

$$\kappa_\chi(3) \coloneqq C_\chi'''(t)|_{t=0} = 8\sum_{j=1}^{m} \lambda_j = 8\text{trace}(\mathbf{RRR}) \coloneqq 8Q(3)$$

and

$$\kappa_\Gamma(3) \coloneqq C_\Gamma'''(t)|_{t=0} = \frac{8}{m}\left(\sum_{j=1}^{m} \lambda_j^2\right)^2 = \frac{8}{m}[\text{trace}(\mathbf{RR})]^2 \coloneqq \frac{8}{m}Q(2)^2,$$

where $Q(k) \coloneqq \sum_{j=1}^{m} \lambda_j^k$. For kurtosis $\kappa(4)$, we find that

$$\kappa_\chi(4) \coloneqq C_\chi''''(t)|_{t=0} = 48\sum_{j=1}^{m} \lambda_j^4 = 48\text{trace}(\mathbf{RRRR}) = 48Q(4)$$

and

$$\kappa_\Gamma(4) \coloneqq C_\Gamma''''(t)|_{t=0} = \frac{48}{m^2}\left(\sum_{j=1}^{m} \lambda_j^2\right)^3 = \frac{48}{m^2}[\text{trace}(\mathbf{RR})]^3 = 48\frac{Q(2)^3}{Q(1)^2}.$$

These results show that the Gamma approximation to $\gamma(\cdot)$ is only matched up to the first two moments, but that higher order moments which include at least skewness and kurtosis do not match between $\Gamma(\alpha, \theta)$ and $\gamma(\cdot)$. Since hypothesis testing of H_0 is generally done at a Type I error that has its quantile in the tail of the null distribution, the quantities of skewness and kurtosis may have a non-negligible influence over the validity of test.

We now show that no Gamma distribution can be found such that its first four moments match those of $\gamma(\cdot)$. For $X \sim \Gamma(\alpha, \theta)$, it is known that $\mathrm{E}(X) = \alpha\theta$, $\mathrm{Var}(X) = \alpha\theta^2$, $\mathrm{Skew}(X) = 2/\sqrt{\alpha}$, and $\mathrm{Kurtosis}(X) = 6/\alpha$. If a Gamma distribution which can match the first four moments of $\gamma(\cdot)$ exists, then we should be able to find α such that the expectation, variance, skewness, and kurtosis of this distribution can match that of $\gamma(\cdot)$. We will show just skewness and kurtosis because it is sufficient to prove the point. It follows that we must find an α which solves

$$\frac{2}{\sqrt{\alpha}} = 8Q(3) \quad \text{and} \quad \frac{6}{\alpha} = 48Q(4).$$

After some algebra we see that we cannot find such an α because $Q(3)^2 Q(4)^{-1} \neq 1/2$. Even in the absence of correlation between SNPs in the gene-based association test, i.e. when $\mathbf{R} = \mathbf{I}$, $Q(3)^2 Q(4)^{-1} = m > 1/2$.

3 Exact test of the gene-based null hypothesis

It is often a target of statisticians to derive the null distribution of a statistic in order to find an exact test of the null hypothesis which it tests. We show in this subsection that finding the exact null distribution of T is not necessary to perform an exact test of its H_0 because of the distributional equivalence between T and S defined above. Recall $T = \mathbf{z}^\top \mathbf{z}$ and $S = \sum_{j=1}^{m} \lambda_j Y_j$ where $Y_j \overset{iid}{\sim} \chi^2(1)$ and λ_j is the jth eigenvalue of the LD matrix $\mathbf{R}$. To more easily demonstrate this test, we will consider just two SNPs with LD matrix $\mathbf{R} \coloneqq r_{12}\mathbf{1}\mathbf{1}^\top + (1 - r_{12})\mathbf{I}$ such that $T = Z_1^2 + Z_2^2$ and $S = \lambda_1 Y_1 + \lambda_2 Y_2$. Since T and S are distributionally equivalent, $P(S > \tau | H_0) = P(T > \tau | H_0)$, and it remains to find τ such that $P(S > \tau | H_0) = \alpha$, our desired Type I error. We can rewrite this statement as $P(T > \tau | H_0) = P(S > \tau | H_0) = P(\lambda_1 Y_1 + \lambda_2 Y_2 > \tau | H_0) \coloneqq P(X_1 + X_2 > \tau | H_0) = \alpha$ where we used the transformation $X_k = \lambda_k Y_k$ for $k = 1,2$. If we can find τ, we only need to compare T to it to perform a controlled test at level α. By the definition $P(X_1 + X_2 > \tau | H_0) = \alpha$, τ defines a null region in the joint space of (X_1, X_2) that is a half-plane of probability mass $1 - \alpha$ in $f_{X_1,X_2}(x_1, x_2)$. We therefore seek an integration of $f_{X_1,X_2}(x_1, x_2)$ over the joint space of (X_1, X_2) to find τ. First, let $F_Y(y)$ represent the cumulative density function (cdf) of $Y \sim \chi^2(1)$ and $f_Y(y)$ be its pdf. We must begin by finding the joint distribution of (X_1, X_2) which is

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{\lambda_1 \lambda_2} f_Y\left(\frac{x_1}{\lambda_1}\right) f_Y\left(\frac{x_2}{\lambda_2}\right), \qquad 0 < x_1 < \infty, 0 < x_2 < \infty\,.$$

The joint pdfs of (X_1, X_2) for $r_{12} \in \{0.0, 0.5, 0.9\}$ are shown in **Figure 1** below. This figure shows that increasing correlation between Z_1 and Z_2 leads to greater differences between λ_1 and λ_2, the eigenvalues of $\mathbf{R}$, which pull the probability mass of $f_{X_1,X_2}(x_1, x_2)$ towards their axes when they are large. We illustrate this property by finding the pdf of the angle ξ of a random vector in the (X_1, X_2) plane from the positive X_1 axis. It can be shown that

$$f_\Xi(\xi) = \frac{\sec^2(\xi)}{\sqrt{\eta}\pi} \frac{1}{\sqrt{\tan(\xi)}} \frac{1}{1 + \tan(\xi)\eta^{-1}}$$

where $\eta = \lambda_1/\lambda_2$. **Figure 2** shows the distribution of ξ under the three scenarios of $r_{12} \in \{0.0, 0.5, 0.9\}$ as before, and these results simply illustrate that a random vector in the space of (X_1, X_2) is more likely to point in the direction of X_1 as r_{12} moves from 0 towards 1. When $r_{12} = 0$, a random vector in this space is equally likely to point towards X_1 as it is to point towards X_2.

Figure 1: Joint distribution of (X_1, X_2) whose convolution produces S, T

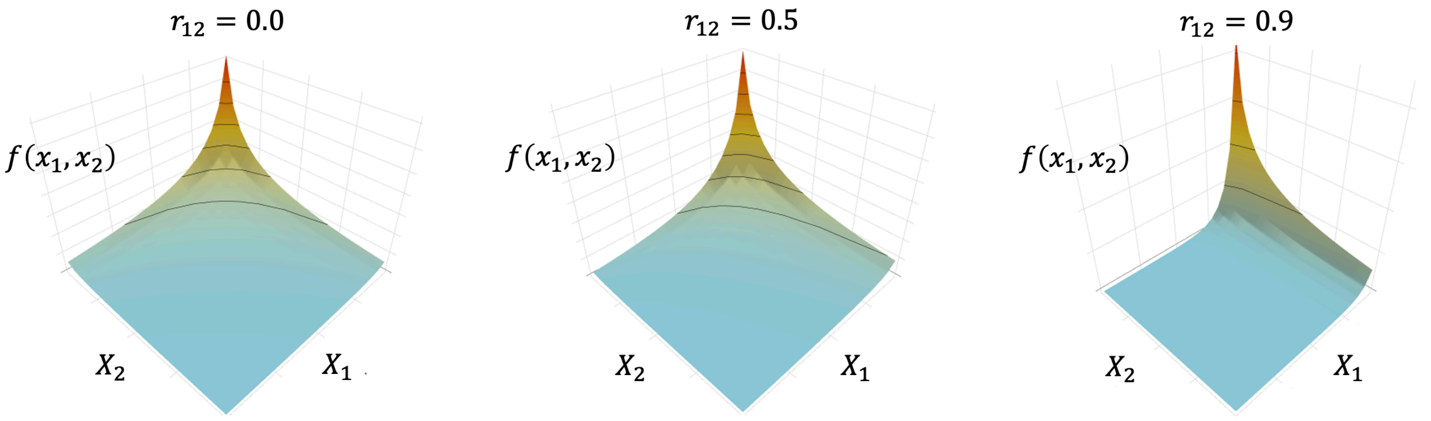
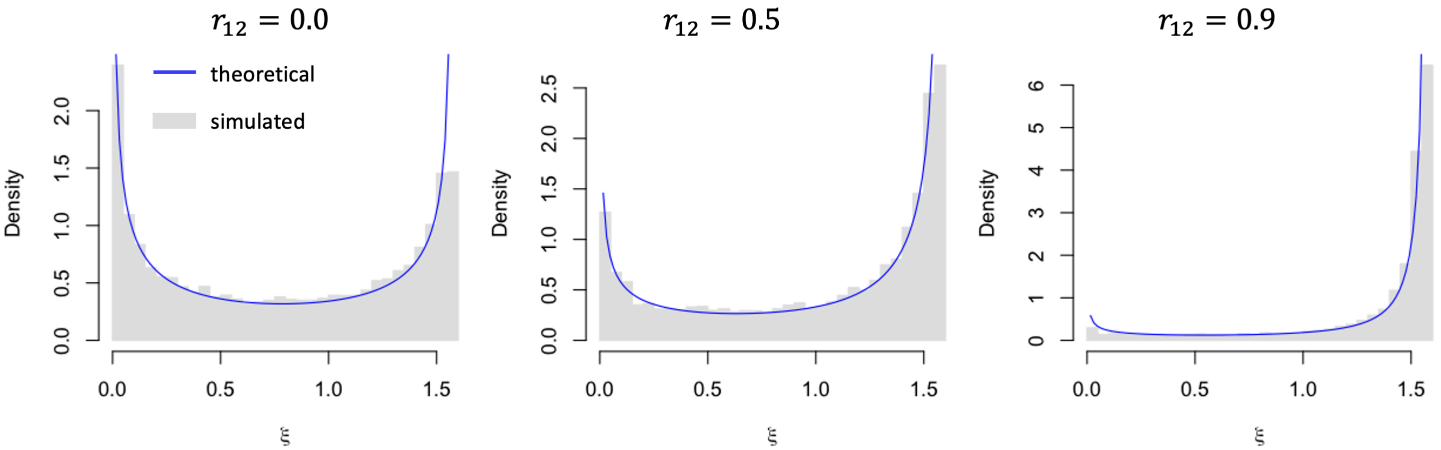


Figure 2: Distribution of radian angle ξ of a random vector in (X_1, X_2) space

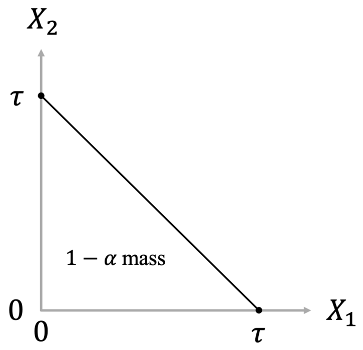


We now turn our attention to finding τ which defines a half-plane in the two-dimensional space of (X_1, X_2) that is the null region of our gene-based association test. An illustration of the τ which we seek is shown in **Figure 3** below. This display shows that the following integration can be used to solve τ:

$$\int_0^{\tau} \int_0^{\tau - x_1} f_{X_1 X_2}(x_1, x_2) dx_2 dx_1 = 1 - \alpha.$$

This integration may be solved analytically, but in higher dimensions, i.e., when we are considering more than just two SNPs in a gene-based association test, the analytical integration may become intractable. As a result, we propose a Monte Carlo sampling scheme to solve the above integral. In the p-dimensional case, we first generate K random observations from the joint distribution $f_{X_1,\ldots,X_p}(x_1, \ldots, x_p)$ to produce K p-length vectors in the set $\mathcal{X} = \{\mathbf{x}_\ell^*\}$. We then define τ as the $(1-\alpha)$th empirical quantile of $\mathbf{1}^\top \mathbf{x}_\ell^*$ and use it to test the gene-based null hypothesis with the observed statistic T.

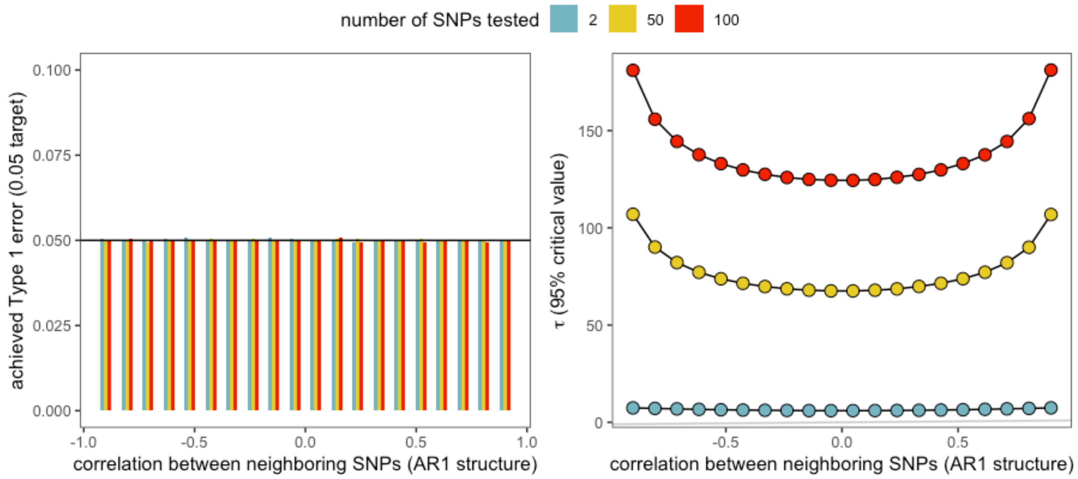
Figure 3: Null region of T as the half-plane in (X_1, X_2) space defined by τ



4 Simulation study

In this subsection we describe a simulation we performed to evaluate the exact gene-based null hypothesis test which is derived from the integration over a p-simplex defined by τ in a general $(X_1, \ldots, X_p)$ space. We generated 1 million independent random samples of T by generating $\mathbf{z}_* \sim N(\mathbf{0}, \mathbf{R})$ multiple times and calculating $T_* = \mathbf{z}_*^\top \mathbf{z}_*$ each time. In this simulation, $\mathbf{R}$ had a first order autoregressive correlation structure with the correlation between neighboring SNPs in the interval $[-0.9, 0.9]$ in steps of approximately 0.95 and the number of tested SNPs in the set $\{2, 50, 100\}$. We set the desired Type I error at 0.05 (5%) and compared our achieved Type I error to this. Our achieved Type I error was defined as the proportion of observations exceeding the simulation setting-specific true τ value, where τ was calculated as the value which solved the integral equality above using the described Monte Carlo approach. Figure 4 presents the results of Type I error and compares our estimated τ values to the true values, which in our simulation was the empirical 95% quantile of the set of T_* observations. These results suggest that the test based on the null region of T constructed in the $(X_1, \ldots, X_p)$ space controls Type I error and therefore could be used to test H_0 at a controlled level after observing T. These results also show that increasing absolute LD between tested SNPs increases the critical threshold τ, as does, unsurprisingly, the number of tested SNPs.

Figure 4: Exact test Type I error simulation results


number of SNPs tested
achieved Type 1 error (0.05 target)
correlation between neighboring SNPs (AR1 structure)
τ (95% critical value)
correlation between neighboring SNPs (AR1 structure)